LA-UR-21-29982

Title: LANL-UW Collaboration on LC-QTOF Datasets – Annual Report

Author(s): Sch neich, Sonia
Cain, Caitlin
Synovec, Robert E.
Freye, Christopher Edward

Intended for: Report

Issued: 2021-10-08

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

LANL-UW Collaboration on LC-QTOF Datasets – Annual Report

September 20, 2021

Sonia Schöneich, Caitlin N. Cain, and Robert E. Synovec

Department of Chemistry, University of Washington, Box 351700, Seattle, WA 98195

Chris Freye

High Explosives Science and Technology, Los Alamos National Laboratory, Los Alamos, NM

Current Work

Prior to the application of non-targeted chemometric techniques (i.e., Fisher ratio analysis), pre-processing steps must be carried out on the data collected using liquid chromatography-quadrupole time of flight (LC-QTOF) mass spectrometry. Data compression is the most important pre-processing step due to the vast number of high-resolution mass channels (*m/z*; 0.0001 Da) collected. Our first approach was to bin the mass spectral dimension to 0.01 Da; however, this approach was computationally expensive and did not preserve the high-resolution data. Therefore, we started using a data compression strategy outlined by Tauler et al., which defines "regions of interest" (ROI) in the LC-QTOF data (Fig. 1). Since high resolution MS data provides an irregular number of measured *m/z* and signal intensity pairs at each retention time, this method searches for ROIs in the raw data that has signals higher than a set threshold and are present in high densities. Those ROIs are then reorganized into a data matrix, which contains the signal intensities for the ROI *m/z* at every time point. These first two steps are repeated for each individual chromatogram in the analysis. Next, to analyze multiple samples simultaneously, the individual ROI data matrices must be augmented. It is important to note that individual ROI matrices can have different ROI *m/z* values; therefore, the augmentation step includes the signal of all common and uncommon ROI *m/z*. The ROI augmentation is first performed for each class and then the two classes are combined to create a super-augmented data matrix. This super-augmented data matrix can then be folded into a three-way array for chemometric analysis. The advantages of the ROI strategy for data compression are that it preserves the mass accuracy of the LC-QTOF data, takes less time to perform on these complex datasets, and requires less computational storage space.

The ROI methodology, outlined in Fig. 1, requires the analyst to define three parameters based upon the experimental design: 1) signal threshold, 2) *m/z* error (i.e., the admissible mass deviation of experimental measurements), and 3) minimum number of retention times to be considered a ROI. All three of these parameters must be defined prior to performing the individual ROI search on the individual chromatograms and then only the *m/z* error must be assigned during the augmentation steps. To study the selection of these parameters, we analyzed the dataset that contained 34 spiked pesticides in soil. Fig. 2 shows the overlaid chromatograms for three spiked pesticides (fludioxonil, paclobutrazol, and triadimenol) at 50 ppb. While all three of these analytes were spiked in at the same concentration, differences in ionization efficiency along with the possibility of ion suppression resulted in the wide range of signal intensities observed. Therefore, while guidance on the selection of these parameters has been given in the literature, a closer examination of these parameters was necessary.

First, the effect of these parameters on the performance of the individual ROI search was studied. The signal threshold ensures that only *m/z* with intensity values larger than the threshold are kept during the data compression. Optimal selection of the signal threshold was shown to be most important for compression of the individual data matrices. To optimize this threshold, we performed the first two steps of the ROI strategy (Fig. 1) on the raw chromatograms for the 50 ppb spiked samples while changing the signal threshold. The final *m/z* kept after performing the ROI search was compared to the list of expected ion fragments for the spiked pesticides to determine the signal threshold that maximizes the number of detectable pesticides while minimizing irrelevant *m/z*. Three different signal thresholds were evaluated: 0.1 % of the max signal intensity (~20000; the lowest threshold suggested by Tauler et al.), 1000, and 250. Six pesticides were found using the lowest signal threshold suggested by Tauler et al. However, the spiked pesticides at their highest concentrations can have intensities below this threshold (Fig. 2), so lower thresholds were evaluated. Five pesticides were discovered in addition to the previous analytes when a threshold of 1000 was used during the ROI search. Likewise, three more pesticides were discovered in addition to the previous ones with a signal threshold of 250. In total, 14 out of 34 pesticides that were spiked in at the highest concentration (50 ppb) were detectable in the compressed data. The signal threshold of 250 was chosen to be appropriate for this dataset because it ensured the maximum number of detectable analytes while keeping the data matrices computationally manageable.

After the data is compressed, an augmented data matrix is created following the same protocol outlined by Tauler et al., first by class and then to include all classes. These data augmentation steps are critical to ensure all samples have the same *m/z* dimension for F-ratios to be calculated on a per *m/z* basis. For the ROI augmentation step, selection of the *m/z* error is important to ensure that signals from the spiked pesticides are not split between multiple *m/z*. Currently, the literature provides much more guidance to optimization of the initial ROI search, however very little guidance is given for selecting appropriate parameters for the augmentation steps. In addition, data sets used in the current literature are not attempting to discover analytes that are present in trace levels, so we needed to optimize parameters to handle the low signal analytes in lower concentration samples. The optimal *m/z* error for the final data augmentation was selected after evaluating the PCA scores plots of all spiked soil classes using the selective *m/z* for the most intense spiked analyte and the top five most intense analytes, which clearly clustered all classes in order of increasing concentration. The optimal *m/z* error determined for the augmentation by class was 0.01 Da. Deviating from this *m/z* error resulted in poor clustering on the PCA scores plot regardless of *m/z* error selected for the super augmentation step. The *m/z* error selected for the super augmentation step was 0.003 Da as it clusters each class correctly by increasing concentration (Fig. 3). This can be further observed by inspecting the PCA scores plot of all spiked classes using only the top hit fludioxonil as well as an overlay of the peak profile for all classes (Fig. 4). It is evident that each class is being clustered well and that we will likely not be able to find hits in the 100 ppt class. Lowering the *m/z* error (0.001 Da) resulted in splitting peak signal among several *m/z*. Using a higher *m/z* error range (0.005 Da) sums too many *m/z* signals per peak and resulted in a couple 25 ppb peak signals to be the same as the 50 ppb and some 10 ppb signals to be the same as some 25 ppb signals, obscuring class differences.

To further investigate the appropriate selection of *m/z* error for the super augmentation step, a selection of 21 *m/z* error values increasing in smaller increments from 0.001 Da to 0.01 Da were used for PCA analysis. For each of the *m/z* error values selected, the degree-of-class separation (DCS) was calculated for several class pairs including 50 ppb and 25 ppb, 25 ppb and 10 ppb, 10 ppb and 100 ppt, 1 ppb and 500 ppt, and 500 ppt and 100 ppt using a representative analyte fludioxonil ($m/z_{\text{Theoretical}}$ 247.0322). For brevity, only results for the first two class pairs (50 ppb and 25 ppb; 25 ppb and 10 ppb) are provided in Fig. 5 in blue. While the DCS was expected to provide a robust metric for determining a range of appropriate *m/z* error that would result in the best representation of the data, DCS should not be considered alone. When low *m/z* error values are selected, such as 0.0013 Da, signal is split among several *m/z* and therefore some DCS values are inflated for this *m/z* error. For example, the DCS calculated for the 50 ppb and 25 ppb class comparison using *m/z* error of 0.0013 Da is 2.97 compared to a DCS of 1.83 using *m/z* error of 0.0017-0.0035 Da. When the *m/z* error 0.0013 Da is selected, the signal is split among several *m/z* so there is signal for the fludioxonil peak ($m/z_{\text{experimental}}$ 247.0331, nearest to theoretical *m/z*) in the 25 ppb class and no signal in the 50 ppb class. On the other hand, when the *m/z* error in the 0.0017-0.0035 Da range is selected, the $m/z_{\text{experimental}}$ 247.0337 nearest to the theoretical *m/z* includes signal across all samples at reasonable concentration ratios that would be expected for known spike levels. Following this observation, the concentration ratios were also factored into the comparisons and are visible in Fig. 5 in orange with the dashed red line representing the true concentration ratio. It was observed that for the *m/z* error range 0.0017-0.0035 Da the same DCS and concentration ratios were calculated for all *m/z* error values (due to the same ROI *m/z* values) with concentration ratios close to true concentration ratios and some of the higher DCS values.

After optimization of the ROI parameters (described above), Fisher ratio (F-ratio) analysis was performed on the spiked pesticide and PBX 9501 datasets. F-ratio analysis is a supervised feature selection technique, which aims to discover class-distinguishing analytes between different classes. Traditionally, F-ratio analysis is defined at the ratio of the between-class variance to the pooled within-class variance. Recently, our group has shown that using the only the smallest within-class variance in the denominator of the F-ratio calculation allows for the discovery of additional class-distinguishing analytes. This calculation, termed the minimum variance optimized (MVO) F-ratio, will be used herein. Along with selecting the appropriate F-ratio metric, four different ways of calculating F-ratio for the datasets were compared (Table 1).

Initially, a pixel-based F-ratio algorithm (approach #1) was used to discover the spiked analytes in the 50 ppb versus neat soil sample comparison followed by a redundant hit removal code to remove false positives. However, a small amount of retention time shifting can be observed in the data, which can cause erroneous results in the hit lists. Additionally, the *m/z* dimension includes over 20,500 *m/z* and a pixel-based approach would calculate ~4.9×10^7 F-ratios prior to the use of thresholds or a redundant hit removal algorithm, which is computationally expensive. Therefore, to account for misalignment in the data and reduce computational time, we tried to develop an F-ratio approach that leveraged the high selectivity of the high-resolution mass spectrometer (approach #2). It was evident from plotting several of the spiked pesticide hits that the *m/z* were pure, so the new F-ratio algorithm (approach #2) finds the peak maxima on a per-*m/z* basis and calculates an average pin location. Peak intensities were then summed across a window of the average peak maxima ± 10 data points, which were then used to calculate the F-ratio and *p*-value. This limits the maximum number of discovered hits to the *m/z* dimension and does not require redundant hit removal based on *m/z*. Preliminary results were able to confidently identify three new hits while discovering all 14 expected hits (lowest hit #213). In addition, this F-ratio method (approach #2) was applied to the PBX9501 and PBX9501_aged samples to determine if there were any obvious class-distinguishing analytes using the same optimized ROI parameters. Using this approach several analytes were discovered that had apparent differences in concentration between classes. While this F-ratio approach (approach #2) worked well for the spiked soil data set with HRMS data, the *m/z* containing class-distinguishing peaks for the PBX9501 comparison were not pure and thus, the algorithm could potentially miss multiple peaks on a given *m/z*.

Next, a tile-based approach (approach #3) was developed to ensure all hits would be discovered without assuming one hit per *m/z*. The first application of a tile-based F-ratio algorithm followed the same process, which has been described in the literature (Marney *et al*. 2013; Parsons *et al*. 2015; Freye *et al*. 2020). Here, the F-ratio was calculated at each *m/z* for each tile and then, the hits were pinned and clustered together. Initial evaluation of this method on the spiked soil samples showed that only a few of the pesticides were discovered (9 pesticides) with different tile sizes (i.e., false positives due to detector fluctuations were much of the hit list). To better discover the pesticides, we switched the order of operations in the original tile-based F-ratio code (approach #3). This new method (approach #4) first creates the tiling scheme, then pins and redraws a tile prior to calculating the F-ratio for the tile. The results of the latter F-ratio method (approach #4) are summarized in Table 2, where 16 pesticides were discovered (lowest hit #1082); however, it is of note that they are discovered lower on the hit list than with the previous F-ratio method (lowest hit # 213). To address this, a redundant hit removal code was integrated in the F-ratio code to remove false positives. Hits that remain in the hit list include what appear to be true positives when plotted, however the *m/z* are not in the range expected for the pesticides (*m/z* over 900) that require further inspection. There may be other true positives in the hit list that are not being identified due to experimental *m/z* that are not like the theoretical *m/z* we expect, which will also be explored further. After this initial evaluation, the F-ratio code (approach #4) was applied to the irradiated explosive samples PBX9501 and PBX9501_aged to obtain preliminary results. Three hits high on the F-ratio hit list are provided in Fig. 6, plotted using the top F-ratio *m/z* and providing the hit number in the top left corner of the panels. Below each overlay plot of the hits a mass spectrum is provided from the pin location provided in the F-ratio hit list. Several of the top intensity *m/z* are labeled in each panel for reference, though were not necessarily used to discover hits by F-ratio. As it is uncertain whether the *m/z* in the mass spectrum are used to discover hits by F-ratio correspond to the molecular ion or adducts, we were not able to confidently identify hits or determine a potential chemical formula at this time. These promising preliminary results suggest there are significant differences between the irradiated explosive samples and would require further investigation.

In summary, we have determined an appropriate pre-processing method to import and compress LC-HRMS data using the ROI method that minimizes computational expense while retaining the high-resolution data prior to F-ratio analysis. Furthermore, the parameters for appropriate application of the three steps of the data compression method for low *S/N* analytes was optimized. Following this optimization, both pixel-based and tile-based F-ratio approaches were evaluated for the discovery of spiked pesticides in soil samples. Using the final tile-based F-ratio method (approach #4), 16/34 spiked pesticides in the soil samples were discovered when comparing the neat and 50 ppb samples. Further investigation would be required to determine what the other hits in the hit list are and to determine if other spiked pesticides were discovered but not identified due to unexpected *m/z*. The same tile-based F-ratio approach (approach #4) was used on the irradiated explosive samples providing promising results of between-class differences. As we are currently not able to identify hits, a future goal is to write an algorithm to match our hit spectra to library spectra to obtain a potential identification as a last step in our workflow. We would also aim to apply the same workflow to the remaining TNT samples purified using different methods to understand better why they behave differently. Lastly, we would like to evaluate this workflow for comprehensive two-dimensional LC-QTOF (LC×LC-QTOF) data. The analysis of LC×LC-QTOF data proves to be challenging not only because of the computational storage of the high-resolution data, but also because of challenges associated with modulating LC data. Mobile phase incompatibilities between the first and second dimension can cause distorted peak shapes, which could pose a challenge for the ROI methodology and further investigations are warranted.

Figures and Tables

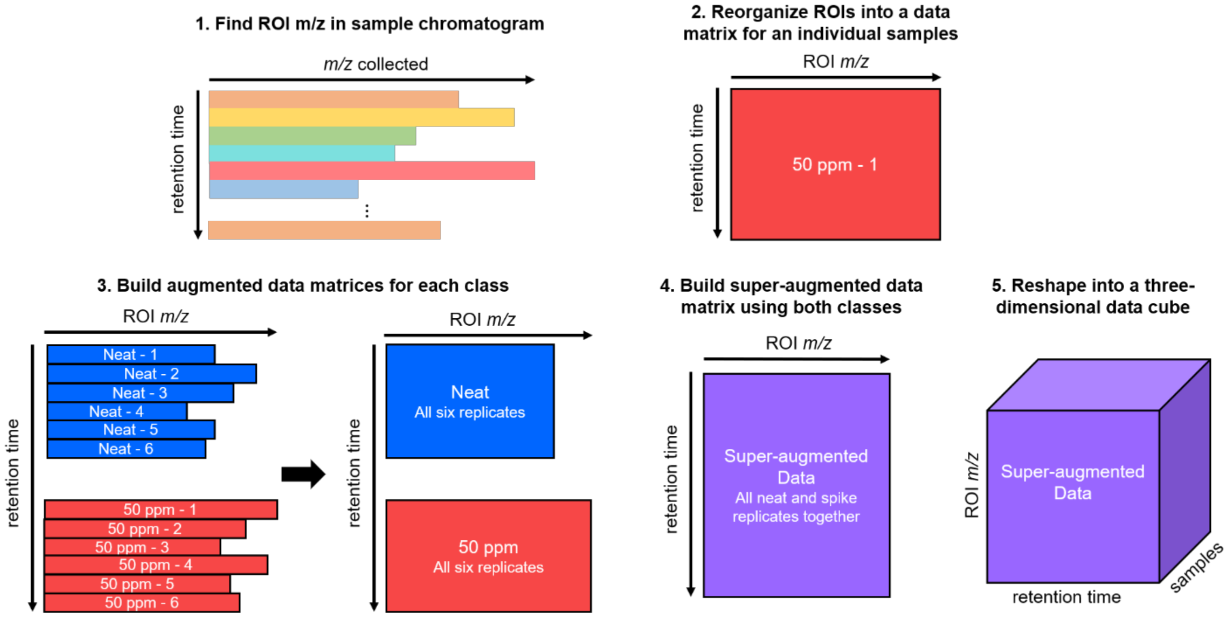

Figure 1. Workflow for the individual and augmented ROI searches.

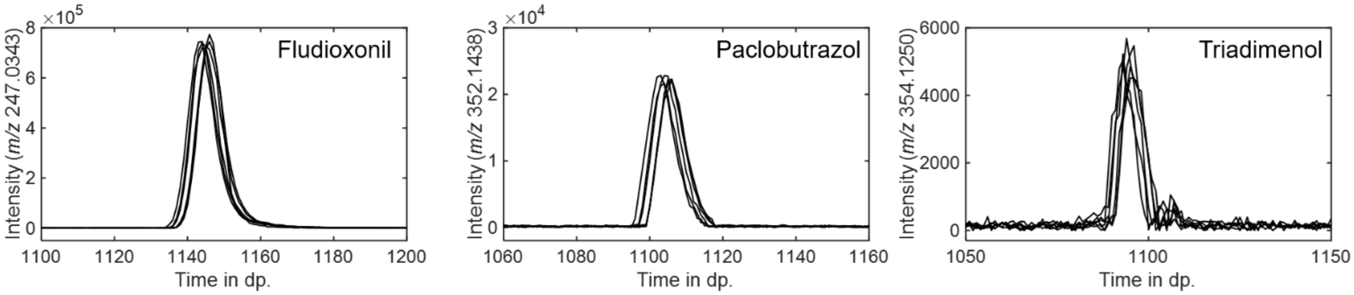

Figure 2. Overlaid analytical ion current chromatograms for fludioxonil, paclobutrazol, and triadimenol.

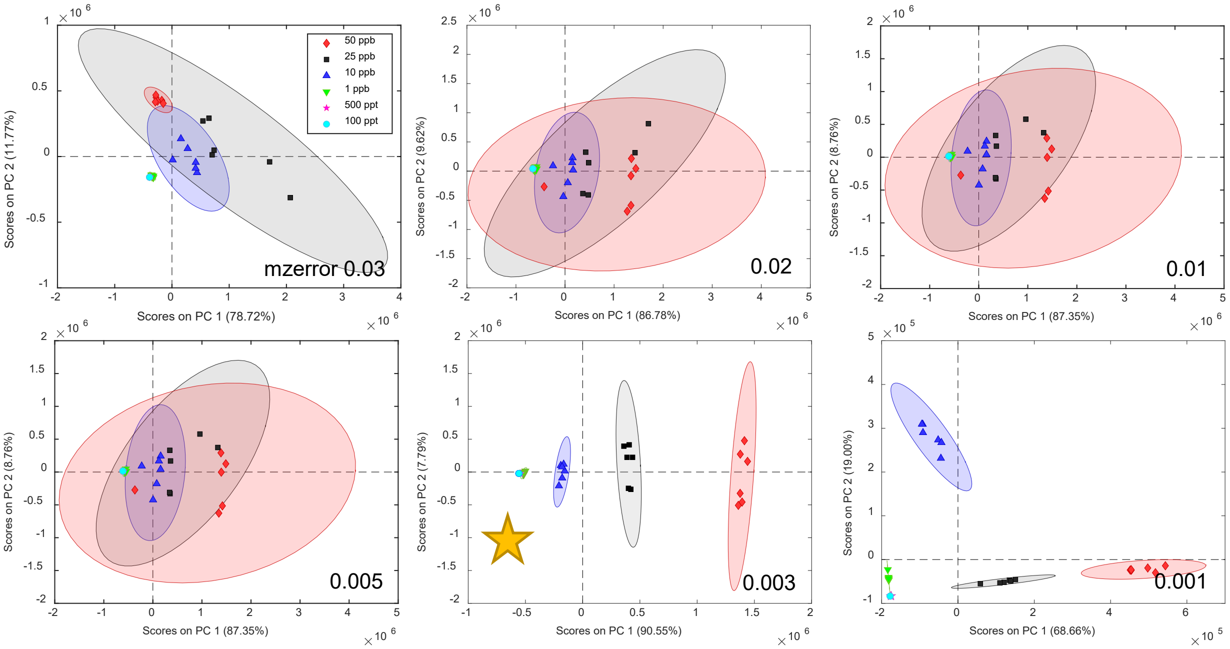

Figure 3. PCA scores plots using peak profiles of top 5 hits evaluating the optimal *m/z* error (specified in lower right corner) to use for super augmentation. The *m/z* error used for the class augmentation step was 0.01 Da with a signal threshold of 0.

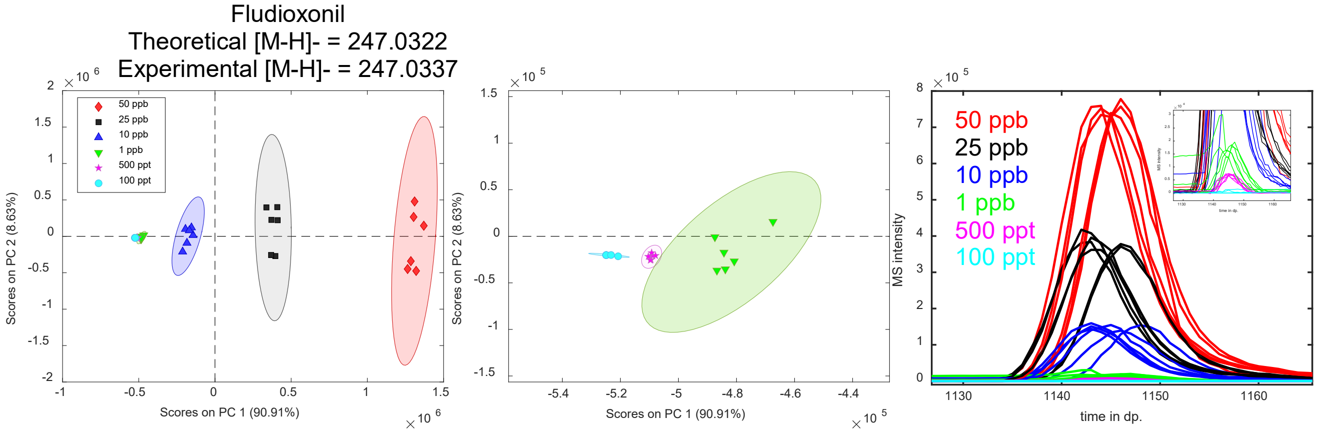

Figure 4. PCA scores plot with zoom in of lower concentrations for top hit fludioxonil and an overlay of all replicates of each class.

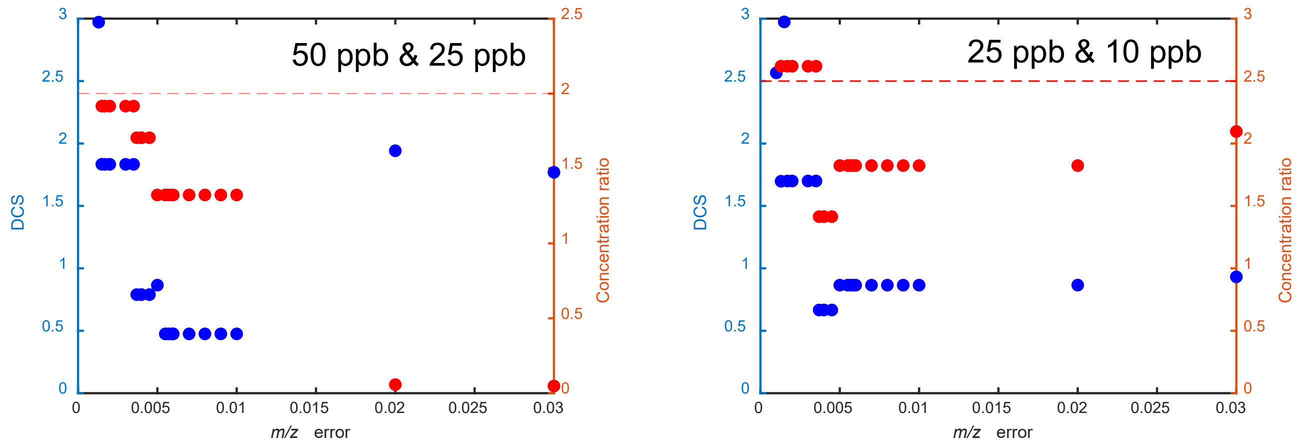

Figure 5. DCS (blue) and concentration ratios (red) calculated for each *m/z* error (0.001, 0.0013, 0.0015, 0.0017, 0.002, 0.003, 0.0035, 0.0037, 0.0040, 0.0045, 0.0050, 0.0055, 0.0057, 0.0059, 0.006, 0.007, 0.008, 0.009, 0.01, 0.02, 0.03 Da) selected in the super augmentation step of the ROI workflow for the 50 ppb and 25 ppb comparison (left) and 25 ppb and 10 ppb comparison (right). The dashed red lines represent the true concentration ratio for each comparison.

Table 1. Summary of the different F-ratio analysis methods utilized.

Approach #	Description
1	Pixel-based F-ratio, where an F-ratio is calculated for every time point on every *m/z.*
2	A peak finder first finds a peak on each *m/z* and draws a window around that peak. The signal within that window is summed together and an F-ratio is calculated.
3	Tile-based F-ratio, where the data is first tiled, and F-ratios are calculated on those tiles. A “pinning and clustering” algorithm is used to remove redundant hits.
4	Also, a tile-based F-ratio method. However, after the data is tiled together, the “pinning and clustering” algorithm is used to find the peak locations. The tile is then redrawn to be centered on those peak locations and F-ratios are calculated.

Table 2. Preliminary F-ratio hit list of the 50 ppb and neat sample comparison generated after optimal parameters were applied. The F-ratio is reported for the top F-ratio *m/z*. The *m/z* reported is the *m/z* used to identify the spiked compound, not necessarily the *m/z* for which the F-ratio is reported. The hit highlighted in green is a new confirmed hit after using the optimized parameters.

Hit no.	ID	t_R	F-ratio	*m/z*	ppm difference
12	fludioxonil	1140	2.23E+09	247.0343	7.5451
14	tricyclazole	1146	1.83E+09	248.0501	0.5257
60	fipronil	1323	1.67E+08	434.9312	0.4096
68	flusilazole	1337	1.33E+08	314.0989	18.7984
76	terbacil	863	1.12E+08	215.0610	7.9946
140	triadimefon	860	3.95E+07	292.0860	0.5583
197	lenacil	1188	2.06E+07	233.1292	1.2721
294	triadimenol	902	8.79E+06	354.1250	6.5241
329	fenarimol	1095	6.42E+06	329.0257	0.8720
365	triflumizole	1155	4.88E+06	344.0826	11.8113
390	flutriafol	1382	4.13E+06	300.0967	4.2477
548	paclobutrazol	971	1.37E+06	352.1438	1.3871
875	procymidone	1106	1.78E+05	282.0104	3.5224
961	myclobutanil	1262	9.63E+04	347.1293	3.6277
1013	tebuconazole	1169	6.59E+04	366.1657	18.3567
1082	hexazinone	1214	3.93E+04	251.1512	0.6706

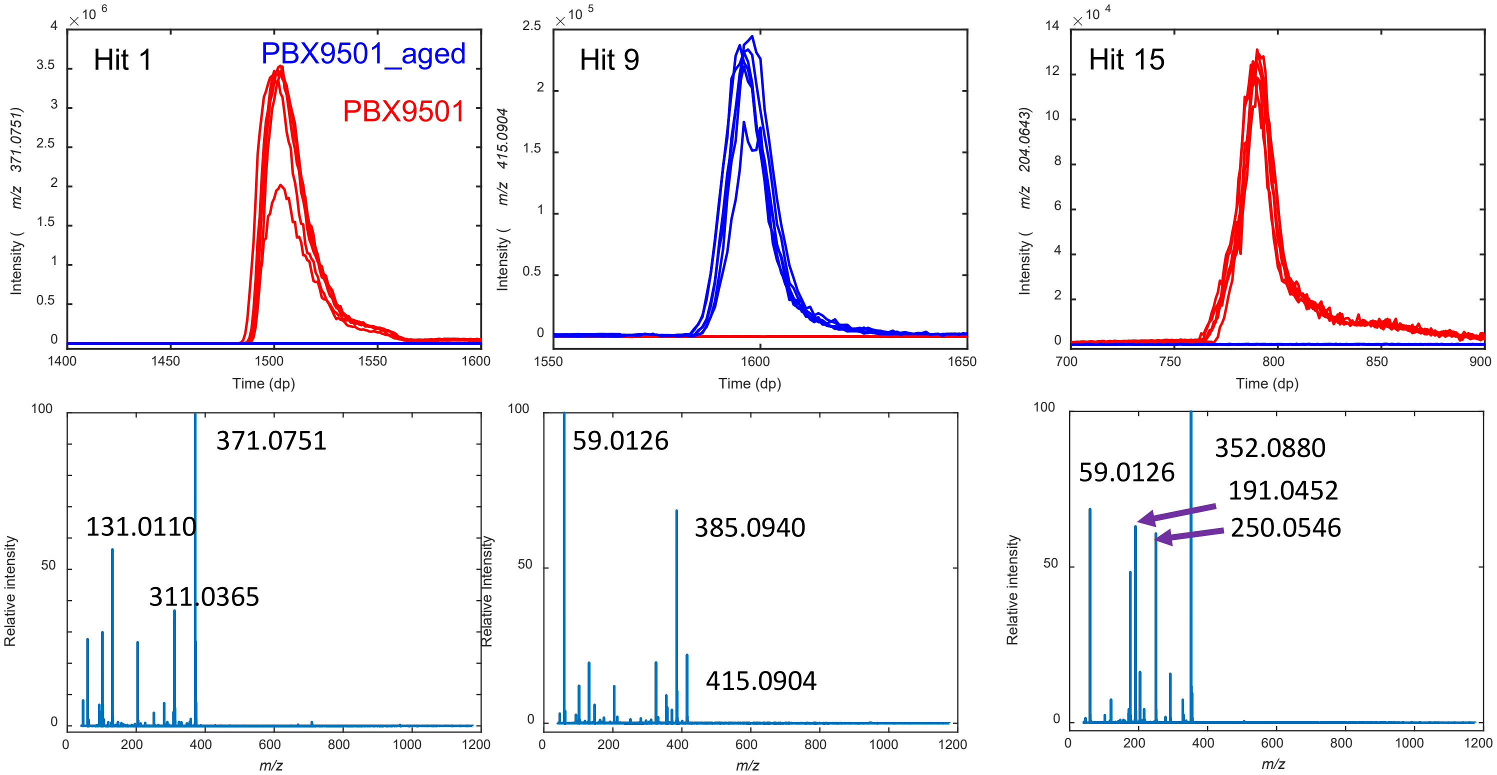

Figure 6. Overlay plots of all replicates of PBX9501 (red) and PBX9501 aged (blue) irradiated explosive samples for three hits discovered by tile-based MVO F-ratio (Top row). Hit numbers are provided in top left corner and *m/z* used to plot is top F-ratio *m/z* for these pin locations. In the bottom row, corresponding mass spectra from the pin location of all hits are provided. Several of the top intensity *m/z* are labeled for reference.